PHIL 7001: Fundamentals of AI, Data, and Algorithms

Week 7
Introduction to Classification

Boris Babic,
HKU 100 Associate Professor of Data Science, Law and Philosophy

Today

Learning goals

- Understand the basic notion of classification and how it forms the core of machine learning.
- Compare regression and classification.
- Analyze a simple classification example.
- Understand the concept of a decision boundary.
- Gain insights into the significance of classification in real world scenarios.

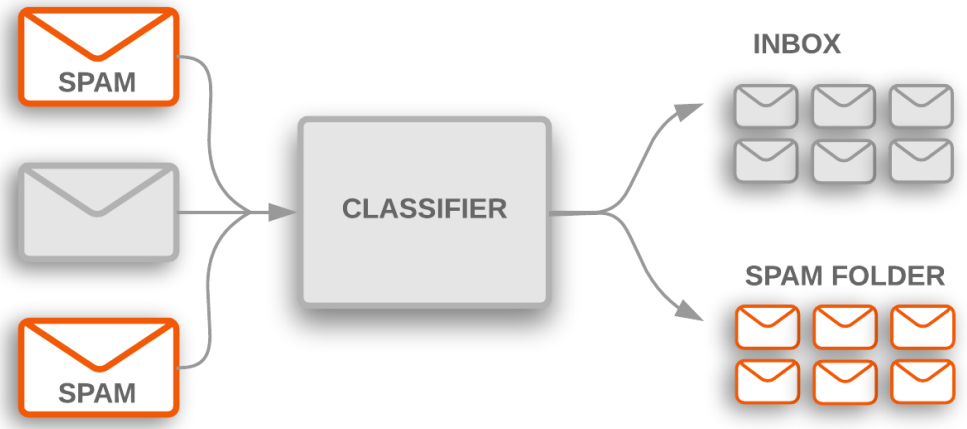
Last week

Review of last week

- Linear regression
- using R to build regression models
- Interpreting model summaries
- Statistical significance and p-values in regression
- Making predictions

What is Classification?

- Classification is a core machine learning task.
- It involves assigning data points to predefined categories or classes.
- Think of it as putting data into labeled bins.
- Simple example: we have collected data on the heights and weights of a group of students. Our goal is to predict who is male and who is female.
- Another example: an algorithm can learn to predict whether a given email is spam or not spam.



Why Classification?

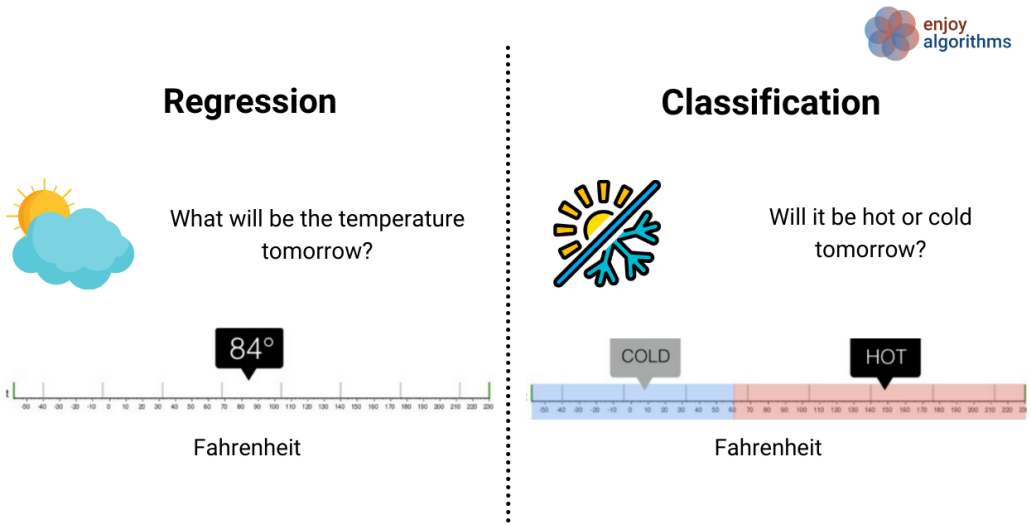
- Classification: a kind of grouping summary
- Easy to (1) interpret
- Easy for (2) making predictions
- Easy for (3) making decisions
- Crucial in many fields: finance, healthcare, marketing, operations and supply chain logistics.

- In classification problems, each entity in some domain can be placed in one of a discrete set of categories: yes/no, friend/foe, good/bad/indifferent, blue/red/green, etc.
- Given a training set of labeled entities, we want to develop a rule for assigning labels to entities in a new (test) set
- Many variations on this theme:
  - Binary Classification
  - Multi-Class Classification
  - Multi-Label Classification
  - Imbalanced Classification

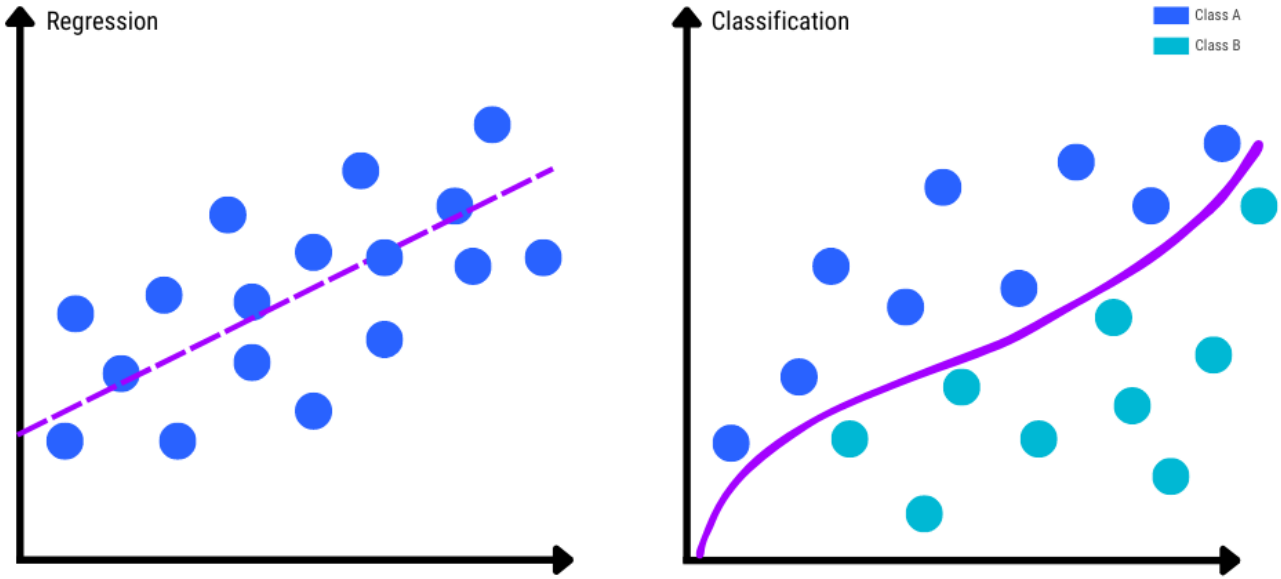
- Each object to be classified is represented as a pair (X, Y):
  - X represents the description of the object (which could be an individual variable, or a vector). As before the X's are called predictors/features.
  - Y is a label/class (categorical, qualitative). As before Y is the response or the output
  - If Y takes two values: two-class or binary classification problem
  - If Y takes more than two values: multi-class classification
- Success or failure of a machine learning classifier often depends on choosing good descriptions of objects
  - The choice of description can also be viewed as a learning problem
  - But good human intuitions are often needed here
- Find a model (decision rule, classifier) for the class variable Y as a function of variable(s) X.
- Goal: previously unseen data points should be assigned a class as accurately as possible.
- Usually, use a training set to build the classifier and test set to evaluate its performance. Then it can be applied in practice to new real world data.

From Regression to Classification

- Recap on Regression:
  - Regression predicts continuous values (e.g., predicting house prices).
  - Fits a line that best represents the relationship between variables.
- Classification:
  - In classification, we predict discrete class labels (e.g., cat or dog, men or women).
  - Focus shifts from estimating quantity to predicting category.



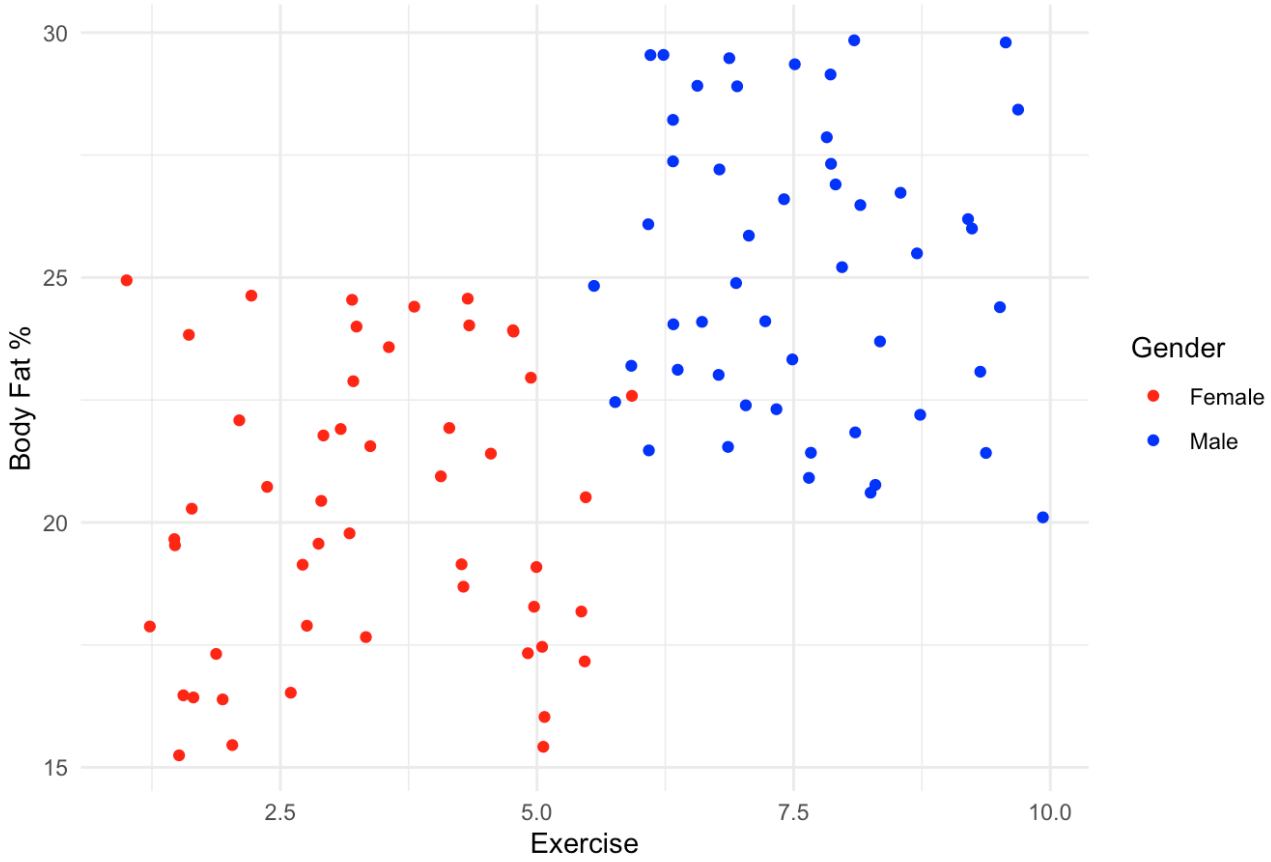
Regression vs. Classification



Simple Classification Example: Predicting Gender

- Recall that in previous lectures, we worked on a scenario where we explored the relationship between body fat percentage and exercise level.
- Now, let's construct a classification model to predict the gender (male or female) based on these two features.
- Binary Classification
  - We're dealing with binary classification (two classes).
  - Male and female are our class labels.

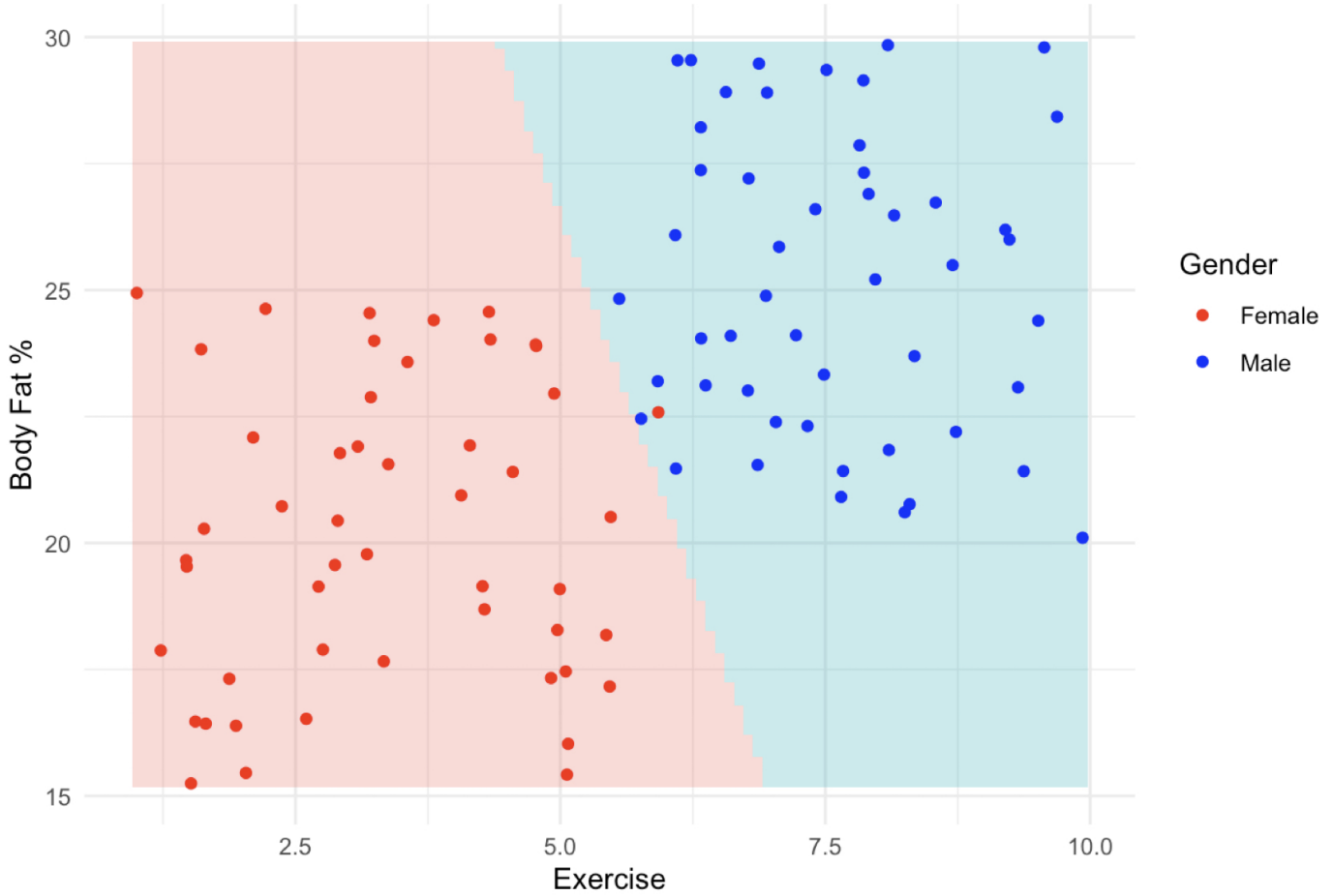
Simple Classification Example: Predicting Gender

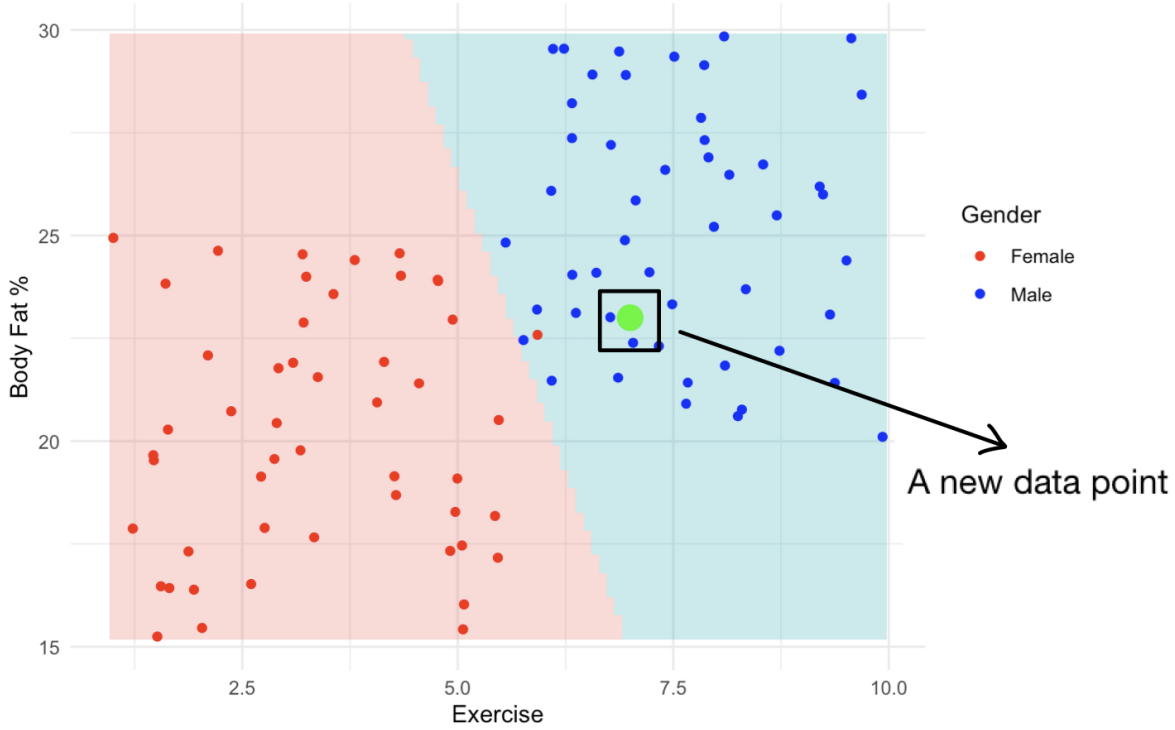


Decision Boundary

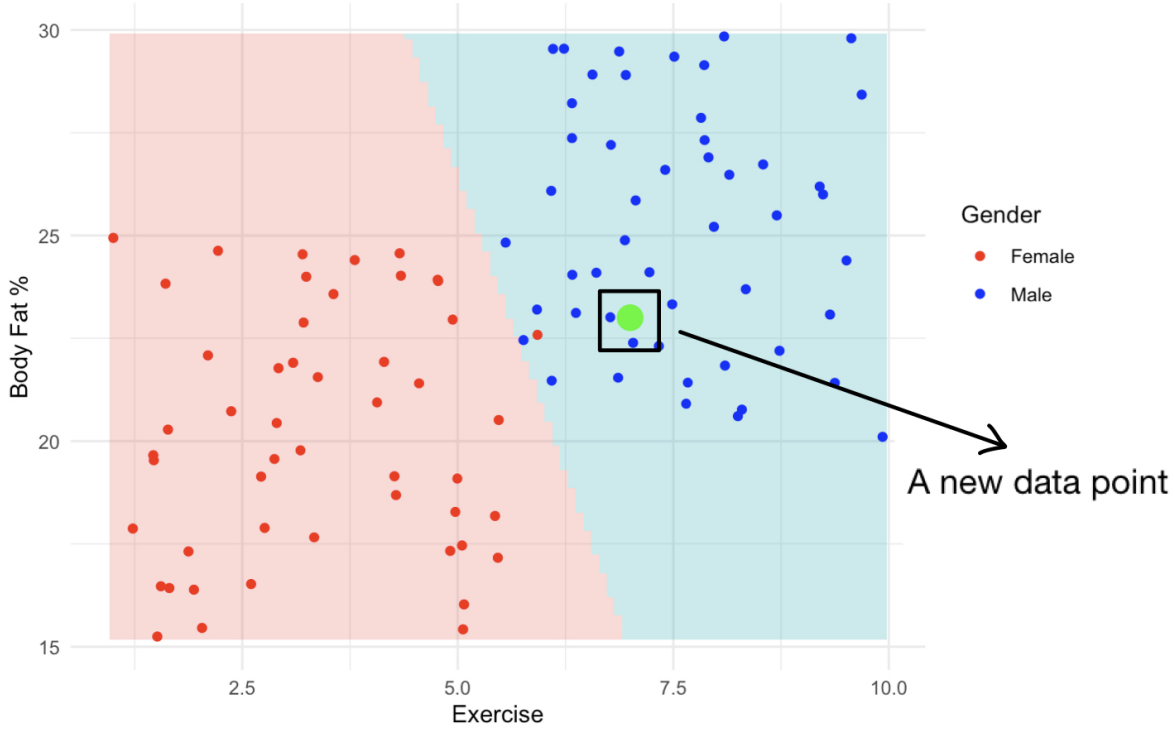
- Understanding Decision Boundary:
  - Decision boundary is a dividing line between classes.
  - For our gender prediction: A line that separates male and female.
- Algorithms find the optimal decision boundary.
- Different classification algorithms lead to different boundary shapes.

Simple Classification Example: Predicting Gender



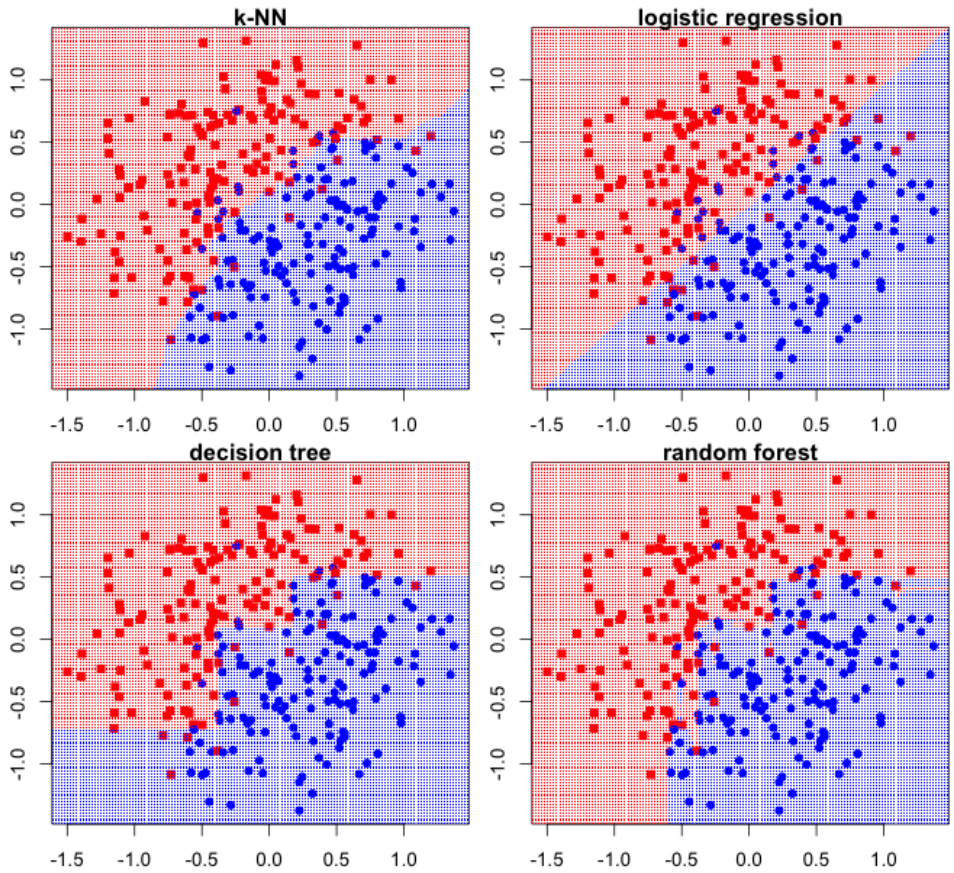


- Now, suppose we know that a student spends around 7 hours per week on exercise and has a body fat percentage of 23%, can you assign a predicted gender label to this student?



- Now, suppose we know that a student spends around 7 hours per week on exercise and has a body fat percentage of 23%, can you assign a predicted gender label to this student?

He will be predicted as Male.



- The decision boundary does not necessarily have to be a straight line.
- Different classification algorithms may lead to different boundary shapes.

Real-World Example - Spam Filter

- Input: email
- Output: spam/ham
- Setup:
  - Get a large collection of example emails, each labeled “spam” or “ham”
  - Note: someone has to hand label all this data
  - Want to learn to predict labels of new, future emails
- Features: The attributes used to make the ham / spam decision
  - Words: FREE!
  - Text Patterns: $dd, CAPS
  - Non-text: Sender In Contacts
  - …

Dear Sir.

First, I must solicit your confidence in this transaction, this is by virture of its nature as being utterly confidencial and top secret. …

TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.

99 MILLION EMAIL ADDRESSES FOR ONLY $99

Ok, Iknow this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.

More Real-World Examples

- Fraud detection (input: account activity, classes: fraud / no fraud)
- Web page spam detection (input: HTML/rendered page, classes: spam/ham)
- Medical diagnosis (input: symptoms, classes: diseases)
- Automatic essay grader (input: document, classes: grades)
- . . . many many more

Today

Learning goals

- Understand the fundamental concepts of classification in machine learning.
- Regression vs. Classification.
- Analyze a simple classification example
- Understand the concept of decision boundary
- Gain insights into the significance of classification in real-world scenarios.

Next lecture

- A particular type of classification: Logistic regression
- Understanding the transition from linear to logistic regression